Ultimate device scaling: intrinsic performance comparisons of carbon-based, InGaAs, and Si field-effect transistors for 5 nm gate length

Mathieu Luisier*, Mark Lundstrom**, Dimitri A. Antoniadis†, and Jeffrey Bokor‡
*ETH Zürich (mluisier@iis.ee.ethz.ch), **Purdue University (lundstro@ecn.purdue.edu),
†MIT (daa@mtl.mit.edu),‡University of California at Berkeley (jbokor@eecs.berkeley.edu)

***Abstract*— We use a single, multi-dimensional, and atomistic quantum transport simulator to investigate how far carbon nanotube, graphene nanoribbon, InGaAs, and Si ultra-thin body and nanowire n-type field-effect transistors can be scaled and to understand the mechanisms that limit their miniaturization. Despite multiple leakage paths, non-planar devices with a multi-gate architecture and an extremely narrow cross section can be expected to still work as good switches, even with a 5 nm gate length, provided that they exhibit a large enough band gap and transport effective mass and that their gate contact can modulate the electrostatic potential of the source and drain extensions to effectively increase the gate length.**

Introduction

By following Moore's scaling law for more than four decades, the size of transistors has been drastically reduced and nowadays gate lengths are below 30 nm [1]. Current approaches such as the triple-gate FinFET recently introduced by Intel [2] are likely to take us to 10-nm gate lengths, but the key question now is: "Can we scale gate lengths below 10 nm?". And if so, how far below? At this scale, the greatest challenges include obtaining (i) good electrostatics of the transistor channel, (ii) minimal source-to-drain tunneling leakage, and (iii) high electron and hole injection velocities that Si might not be able to provide [3].

Many different device designs and material combinations are currently being investigated. Among the most promising structures are carbon nanotubes (CNT) [4], graphene nanoribbons (GNR) [5], III-V or Si ultra-thin-body (UTB) and nanowire (NW) [6-8]. Numerical simulation capabilities have advanced enough so that they can help shed light on the scalability and the strengths and weaknesses of each ultimate device candidate. Our goal here is not to give absolute numbers and definitive answers, but, by examining the intrinsic characteristics of ultra-scaled transistors, to see whether such devices can still operate as good logic switches in the sub-10-nm range, what are the limiting factors, and how their designs could be improved to circumvent the inherent problems.

Hence, for the first time, the ballistic performances of highly idealized n-type gate-all-around (GAA) CNT, single- (SG) and double-gate (DG) armchair GNR, Ω-gated strained Si and $In_{0.75}Ga_{0.25}As$ FETs, and DG strained-Si UTB FETs, all with a gate length down to L_g=5 nm, are compared using the same quantum transport simulator and set of approximations [9]. A similar study was performed before for III-V vs Si devices with L_g=5 nm, but in the effective mass approximation [10]. We show here that (i) by carefully engineering the gate-induced potential barrier, good sub-threshold operation can be obtained and (ii) small diameter Si and III-V NWs as well as CNT FETs exhibit about the same performance when their bandgaps are the same.

Approach

As device analysis tool, we employ a multi-dimensional, atomistic, full-band quantum transport simulator based on the nearest-neighbor tight-binding model ($sp^3d^5s^*$ for strained-Si [11] and $In_{0.75}Ga_{0.25}As$ [12] and single-p_z orbital model for carbon), a computationally efficient wave function approach equivalent to the popular NEGF formalism [9], and a 3-D finite element grid for the Poisson equation. Quantum transport is solved in the ballistic limit of the UTB, NW, AGNR, and CNT FETs.

Results

The idealized devices considered in this work are described in Fig. 1. At L_g=5 nm, several leakage mechanisms limit the sub-threshold swing (SS) of these transistors (Fig. 2): intra-band (or source-to-drain) tunneling (IBT), band-to-band tunneling from the conduction band (CB) of the source into the CB of the drain through the valence band (VB) of the channel (BTBT1, V_{gs}-dependent), hole-induced barrier lowering (HIBL), and BTBT from the source VB into the drain CB (BTBT2, V_{ds}-dependent).

The leakage labeled BTBT2 in Fig. 2 can be avoided if the channel band gap E_g is larger than the drain-to-source voltage V_{ds}, assuming no electron-phonon scattering. At V_{ds}=0.5 V, as chosen here to reduce power consumption and heat generation, CNT and AGNR must be scaled to d=1.49 nm and w=2.09 nm, respectively, to match this condition (E_g=0.563 eV as obtained from the simple p_z-orbital model). There is experimental evidence [13] that the actual CNT quasiparticle band gap for a given diameter may be larger than predicted by the model used here, in which case the results are still applicable, but for the CNT diameter that actually corresponds to the band gap found in our calculations. Still, all the other leakage mechanisms remain active and increase the SS of the considered carbon-based FETs above 200 mV/dec (Fig. 3), making them impracticable for logic applications. It is also found that the SG, DG, and GAA structures exhibit different transfer characteristics, indicating that the electrostatics depend on the gate configuration, even in monolayer structures with very small EOT.

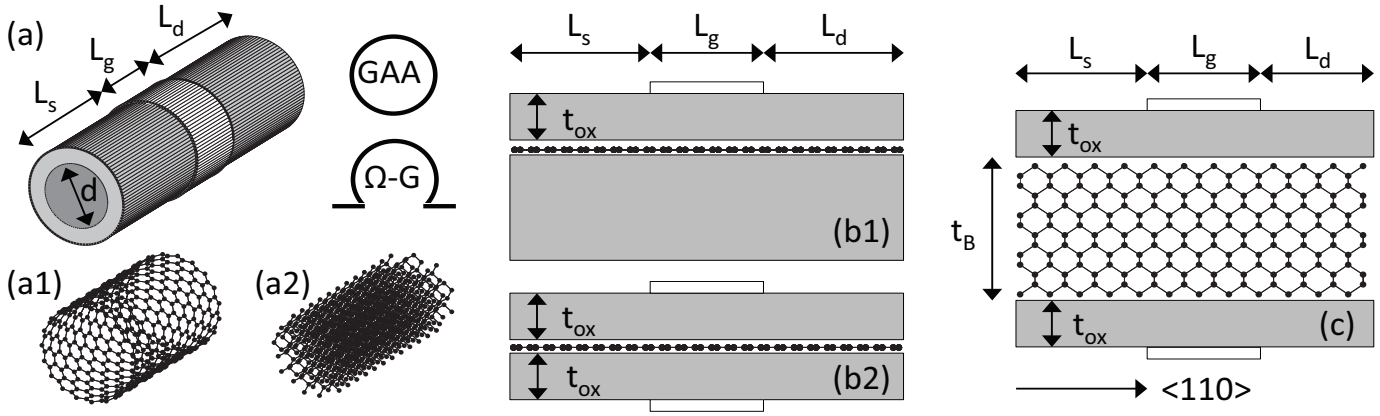

Fig. 1. Schematic view of the different n-type field-effect transistors considered in this work. (a) 3-D structures including (a.1) gate-all-around single-wall carbon nanotubes with diameters d=0.63 nm (doping density N_D=0.8 atoms per nm, band gap E_g=1.408 eV), d=1.02 nm (N_D=0.5/nm, E_g=0.817 eV), and d=1.49 nm (N_D=0.8/nm, E_g=0.563 eV) and (a.2) Ω-gated nanowires (gate covers 75% of the nanowire perimeter) made of strained Si (d=3 nm, N_D=1e20 cm^{-3}, E_g=1.404 eV, transport along the <110> crystal axis, 1% tensile uniaxial strain along <110>) and $In_{0.75}Ga_{0.25}As$ (d=3nm, N_D=7.5e25 cm^{-3}, E_g=1.378 eV, transport along the <100> crystal axis). (b) Single- and double-gate armchair graphene nanoribbon with w=2.09 nm, N_D=0.4/nm, and E_g=0.563 eV. (c) Double-gate Si ultra-thin-body with t_B=3nm, E_g= 1.228 eV, N_D=1e20 cm^{-3}, transport along <110>, confinement along (100), and 1% tensile uniaxial strain along <110>. This strain value corresponds to an applied stress of 1.7 GPa, more than what can be reached nowadays. All the structures have a gate length L_g=5 nm, source and drain extensions L_s=L_d=20 nm, a EOT=0.64 nm made of t_{ox}=3.3 nm HfO_2 with ϵ_r=20, unless stated differently, an abrupt source and drain doping profile aligned with the gate edges, and an applied source-to-drain voltage V_{ds}=0.5 V. The CNT and NW currents are normalized by their diameter, the GNR currents by their width. All the simulations are performed at room temperature.

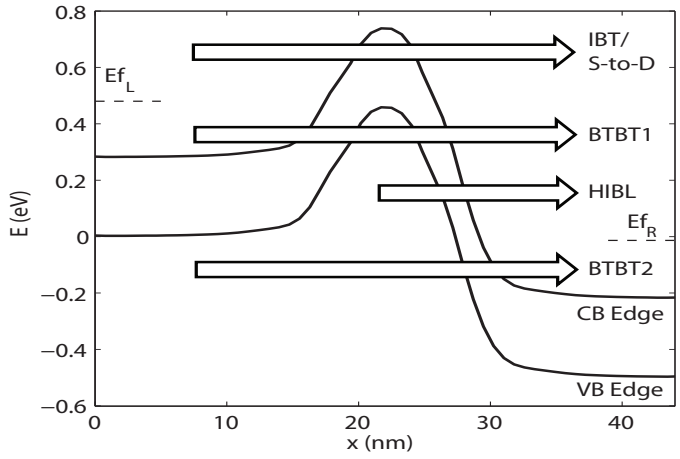

Fig. 2. Illustration of the leakage mechanisms occurring in ultrashort devices. Intra-band or source-to-drain tunneling, hole-induced barrier lowering (HIBL), and band-to-band tunneling in the region 1 (BTBT1) and 2 (BTBT2) are automatically included in ballistic simulations. An additional BTBT path is possible where HIBL occurs if electron-phonon scattering is present.

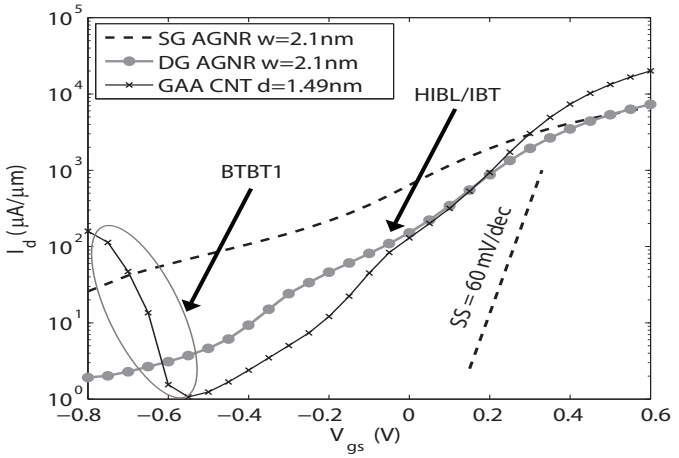

Fig. 3. Transfer characteristics I_d-V_{gs} at V_{ds}=0.5 V of a gate-all-around carbon nanotube (GAA CNT) FET with d=1.49 nm, a single-gate (SG) armchair graphene nanoribbon (AGNR) FET with w=2.09 nm, and a double-gate (DG) armchair graphene nanoribbon (AGNR) FET with w=2.09 nm. The dashed line shows the 60 mV/dec limit of SS in FETs at room temperature.

To improve the SS of ultra-scaled transistors and reduce their tunneling leakage, the shape of their gate-controlled potential barrier should be optimized. As an illustration, GAA CNT FETs with d=1.49 nm and different dielectric configurations (HfO_2, HfO_2+SiO_2 spacers, and SiO_2) are simulated and their conduction band edge and spectral current are reported in Fig. 4. Due to the gate fringing capacitance acting on the source and drain extensions, the effective gate length can be larger than the physical length. With pure HfO_2 dielectric layers, the fringing fields are the largest, the effective gate length increases, and the sub-threshold slope becomes steeper, as shown in Fig. 5, for both CNT and Si NW FETs. As a drawback of strong fringing fields, the source and drain regions cannot scale below ($\geq$20 nm) to maintain a high charge density in the contacts. Note that gate underlap doping produces the same effect.

The leakage mechanisms depend on the bandstructure properties of the devices too, especially IBT. To provide some insight into this issue, the transmission probability $T(E)$ through a potential barrier similar to that observed in transistors with L_g=5 nm is reported in Fig. 6 for four different structures. The accompanying complex bandstructures explain

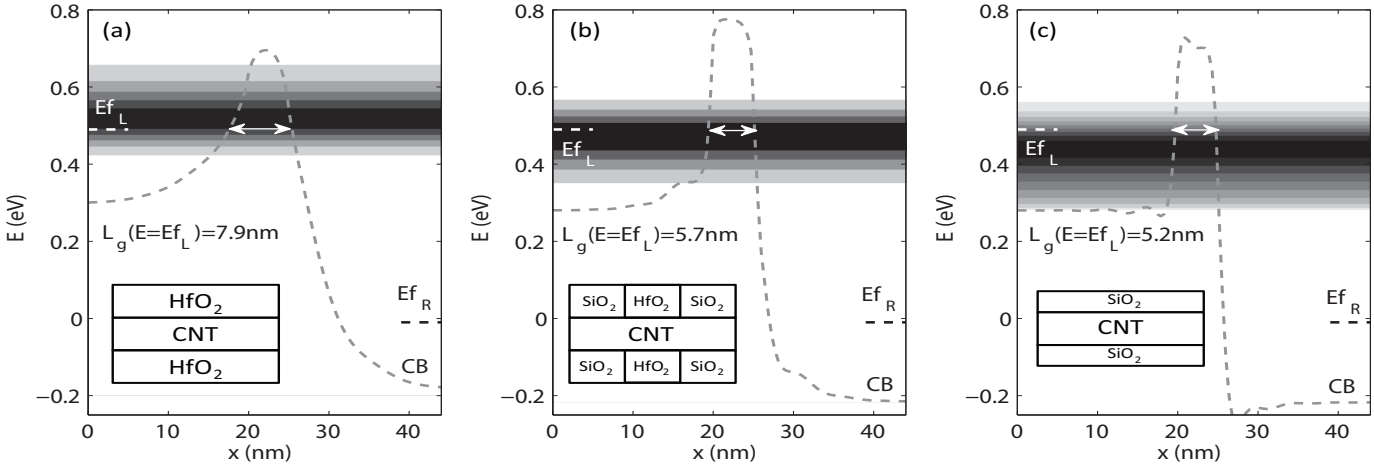

Fig. 4. Spectral current distribution in a GAA CNT FET with d=1.49 nm and EOT=0.64 nm at V_{ds}=0.5 V and V_{gs}=-0.15 V. Dark regions indicate high current concentrations. The conduction band edge (CB) as well as the position of the source (Ef_L) and drain (Ef_R) Fermi levels are also given. (a) The GAA dielectric is made of a 3.3 nm HfO_2 layer all along the device (45 nm), (b) the 5 nm channel is surrounded by a 3.3 nm HfO_2 layer, the 20 nm source and drain extensions by a 3.3 nm SiO_2 layer, and (c) for comparison purpose, the entire device is surrounded by a 0.64 nm SiO_2 layer. The resulting effective gate length at the source Fermi energy is reported as well as the dielectric configuration. The same EOT, defined as EOT=$t_{ox} \cdot \epsilon_{Si}/\epsilon_{ox}$, is used in each case.

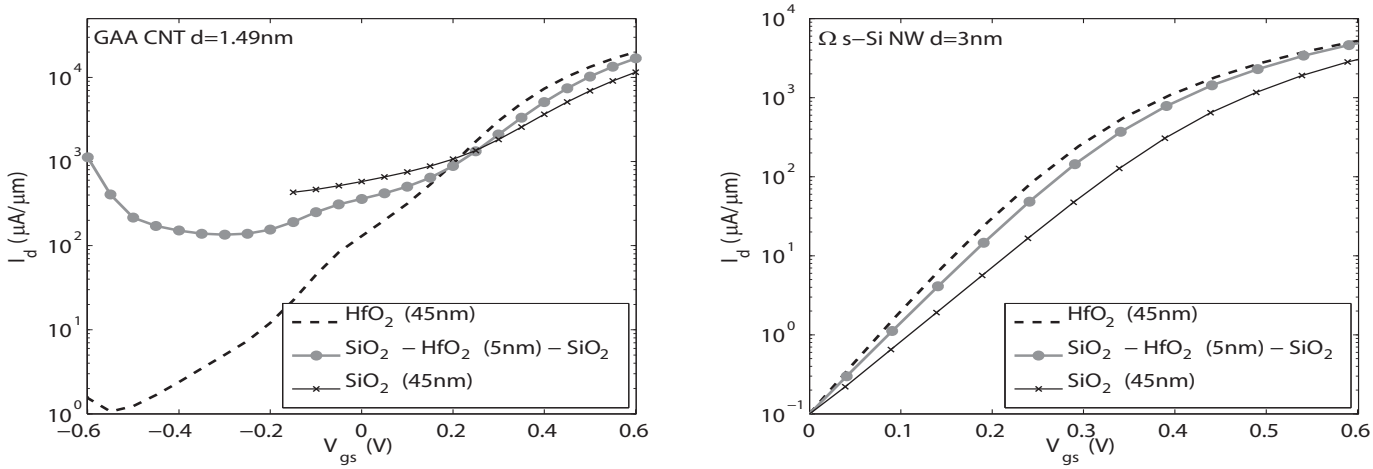

Fig. 5. Transfer characteristics I_d-V_{gs} at V_{ds}=0.5 V of a GAA CNT FET with d=1.49 nm (left) and of an Ω-gated strained-Si nanowire FET with d=3nm. The three same dielectric configurations as in Fig. 4 are used: HfO_2 all-along the device (dashed line), HfO_2 around the 5 nm channel only with 20 nm SiO_2 spacers (solid line with squares), and SiO_2 all-along the device (solid lines with crosses).

why $T(E)$ is material-dependent: a smaller imaginary wave vector $\kappa(E)$ leads to lower wave function attenuation for electrons tunneling through a potential barrier. Such a calculation shows that having a large band gap is not only important to reduce BTBT and HIBL, but also S-to-D tunneling, the value of $\kappa(E)$ between the conduction band edge and the top of the potential barrier depending on E_g. It also reveals that a low transport effective mass, which is good for large ON-currents and high injection velocities becomes a handicap in ultra-short devices due to S-to-D tunneling.

Finally, the electrostatic and bandstructure properties are combined to investigate the performance of five different devices with L_g=5 nm: a GAA CNT with d=0.63 nm, another with d=1 nm, an Ω-gated strained Si and $In_{0.75}Ga_{0.25}As$ NW (d=3 nm), and a DG strained-Si UTB FET. The intrinsic transfer characteristics at V_{ds}=0.5 V are shown in Fig. 7. The CNT with d=1 nm and the DG UTB suffer from a large SS, but the 3 other devices with E_g ∼1.4 eV appear very promising. It is worthwhile noting that the SS of ultra-scaled transistors deteriorates very rapidly for $L_g \leq$8 nm, suggesting that FinFETs might properly work down to this gate length.

Conclusion

We have shown that non-planar devices with $E_g >$1 eV, an effective gate longer than its physical footprint, and ultra-small diameters could provide good intrinsic performance, even at 5 nm gate length. The shape of the potential barrier and IBT have been identified as key factors to design devices with sub-10-nm gate lengths. A sufficiently high band gap is necessary to minimize both drain (BTBT) and source-to-drain (IBT) leakages. As future work, p-type devices and other crystal orientations should be investigated.

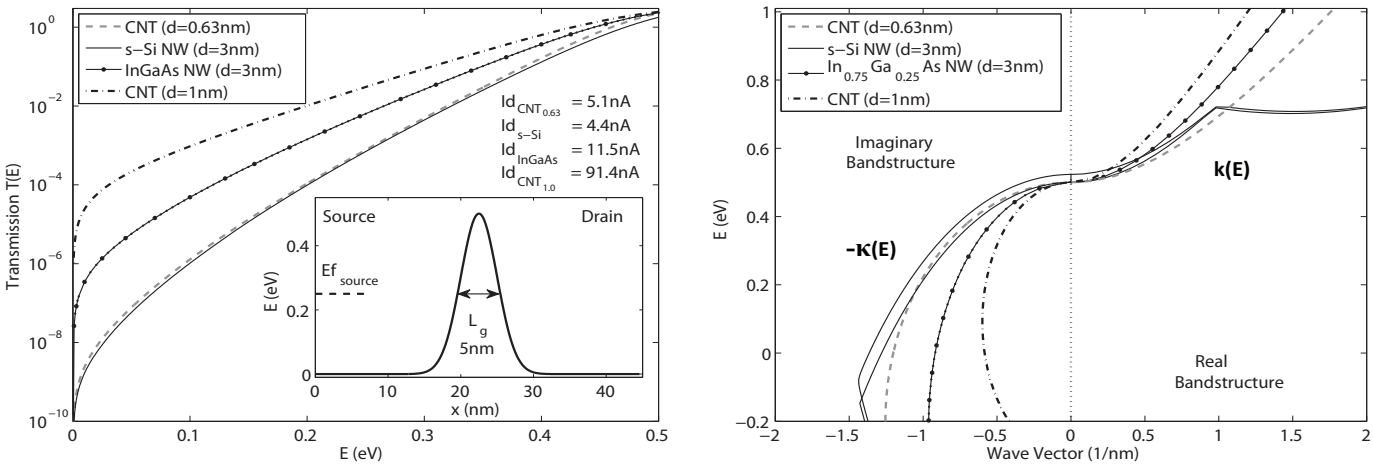

Fig. 6. Analysis of the tunneling properties of a CNT with d=0.63 nm and E_g=1.408 eV, a strained-Si nanowire with d=3 nm, E_g=1.404 eV, and transport along the <110> crystal axis, an $In_{0.75}Ga_{0.25}As$ nanowire with d=3 nm, E_g=1.378 eV, and transport along the <100> axis, and a CNT with d=1 nm and E_g=0.817 eV. (left) Transmission through a 0.5 eV high Gaussian potential barrier with a width of 5 nm at E=0.25 eV (see inset) for the smaller CNT (dashed line), s-Si NW (solid line), InGaAs NW (points), and larger CNT (dashed-dotted line) described above. This mimics intra-band (or source-to-drain) tunneling in an ultra-short FET. By using the same potential profile for the 4 cases, electrostatics effects can be removed and only the bandstructure properties can be investigated. As an indication, the ballistic current flowing through the Gaussian potential barrier is also given with a source Fermi level set to Ef_{source}=0.25 eV and empty states on the drain side. (right) Real and imaginary bandstructure of the same d=0.63 nm CNT (dashed line), s-Si NW (solid line), InGaAs NW (points), and d=1 nm CNT (dashed-dotted line) as before. For convenience, the minimum of the conduction band of all these materials is moved to E=0.5 eV, corresponding to the maximum of the Gaussian potential barrier. The value of the imaginary wavevector $\kappa(E)$ at a given energy E determines the attenuation rate of the wave function $\psi(x, E)$ tunneling through a potential barrier $\psi(x, E) \sim exp(-\kappa(E) \cdot x)$.

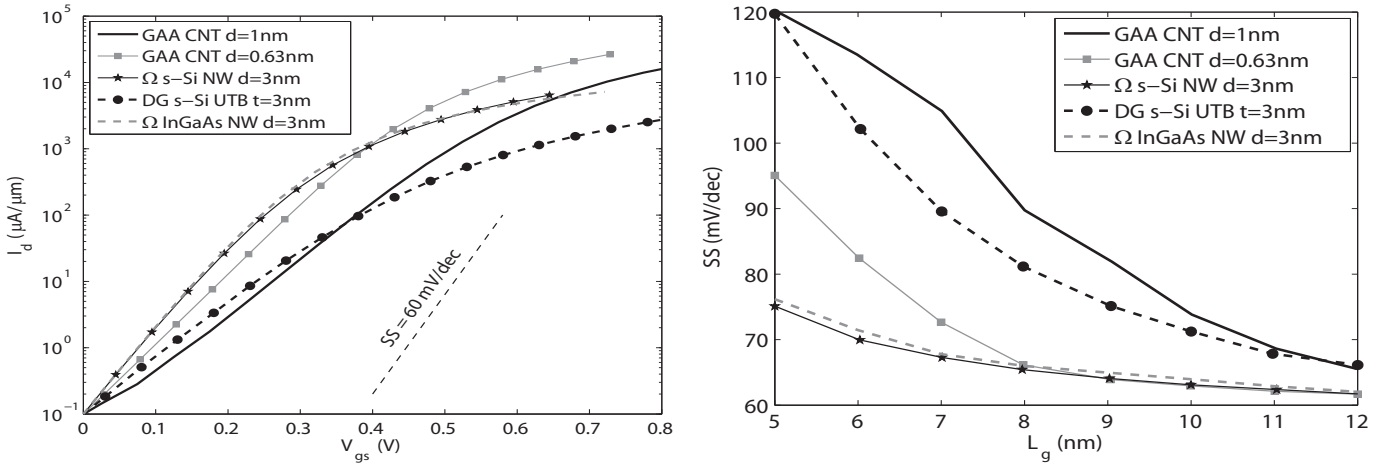

Fig. 7. (left) Transfer characteristics I_d-V_{gs} at V_{ds}=0.5 V of a GAA CNT FET with d=1 nm and E_g=0.817 eV (solid line), a GAA CNT FET with d=0.63 nm and E_g=1.408 eV (solid line with squares), an Ω-gated strained-Si nanowire FET with d=3nm and E_g=1.404 eV (solid line with stars), a double-gate strained-Si UTB FET with t_B=3nm and E_g=1.228 eV (dashed line with circles), and an Ω-gated $In_{0.75}Ga_{0.25}As$ nanowire FET with d=3nm and E_g=1.378 eV (dashed line). The 60 mV/dec slope is indicated for clarity. (right) Sub-threshold slope of the same FETs as before as function of their gate length $5 \leq L_g \leq 12$ nm.

Acknowledgment

This work was supported by the MSD Focus Center, one of six research centers funded under the Focus Center Research Program (FCRP), a Semiconductor Research Corporation entity and by NSF through XSEDE resources provided by the National Institute for Computational Sciences (NICS). This research also used resources of the National Center for Computational Sciences (NCCS) at ORNL, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

[1] P. Packan et al., IEDM Tech. Dig. **2009**, paper 28.4 (2009). [2] http://www.intel.com/technology/architecture-silicon/22nm/ [3] A. Khakifirooz et al. , IEEE Trans. Elec. Dev. **55**, 1391 (2008). [4] A. Javey et al., Nature **424**, 654-657 (2003). [5] X. Wang et al., Phys. Rev. Lett. **100**, 206803 (2008). [6] S. D. Suk et al., IEDM Tech. Dig. **2007**, 891 (2007). [7] D.-H. Kim et al., IEEE Trans. Elec. Dev. **55**, 2546 (2008). [8] A. C. Ford et al., Nano Letters **9**, 360 (2009). [9] M. Luisier et al., Phys. Rev. B **74**, 205323 (2006). [10] K. Cantley, H. Pal, Y. Liu, S. Ahmed, and M. Lundstrom, IEDM Tech. Dig. **2007**, 133 (2007). [11] T. Boykin et al., Phys. Rev. B **81**, 125202 (2009). [12] M. Luisier et al., IEEE SISPAD **2009**, 67 (2009). [13] A. Malapanis et al., Nano Lett. **11**, 1946 (2011).